Available online at www.sciencedirect.com

SCIENCE DIRECT

Biochimica et Biophysica Acta 1762 (2006) 17 – 28

BIOCHIMICA ET BIOPHYSICA ACTA BBA

http://www.elsevier.com/locate/bba

Review

The importance and identification of regulatory polymorphisms and their mechanisms of action

Paul R. Buckland *

Department of Psychological Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

Received 5 September 2005; received in revised form 11 October 2005; accepted 11 October 2005
Available online 28 October 2005

Abstract

The search for the genetic variations underlying all human phenotypes is in its infancy but must be one of the long term goals of the scientific community. There is evidence that most, if not all human phenotypes, including illnesses are influenced by the genetic makeup of the individual. There are an estimated 11 million human genetic polymorphisms with a minor allele frequency >1% and possibly many times that number of rare sequence variants. The proportion of these sequence variants which have any functional effect is unknown but it is likely that the majority of those which influence illness lie outside of the amino acid coding regions of genes, and affect the regulation of gene expression—these are called rSNPs. Recent research suggests that about 50% of genes have one or more common rSNPs associated with them and probably most if not all genes have an rSNP within the human population. In the long term, determining which polymorphisms are potentially functional must be done bio-informatically using algorithms based upon experimental data. However, at the current time, the limited data that has been obtained does not allow the creation of such an algorithm. In vitro studies suggest that a large proportion of rSNPs lie within the core and proximal promoter regions of genes but it is not clear how the majority of these influence transcription, as they do not appear to be within any known transcription factor binding sites. However, promoter regions possess a number of sequence-dependent characteristics which make them distinct from the rest of the genome, namely stability, curvature and flexibility. Subtle changes to these features may underlie the mechanisms by which many polymorphisms exert their function.

Keywords: Genetic variation; Allelic expression; rSNP; DNA curvature; DNA flexibility

1. Introduction

The identification of sequence variation or mutations which cause mendelian or single gene disorders has been ongoing since the early 1980's and the pace of discovery has accelerated rapidly. At the same time, the recognition of the role played by genetic variation in a range of phenotypes and disease has also grown. While the number of illnesses known to have a genetic component runs into the thousands, the influence of genetic variance on many other phenotypes has been demonstrated by twin studies or similar approaches [1].

It is now becoming clear that the pathogenesis or etiology of the vast majority of human diseases is influenced to some extent by genetic variation. This includes illness whose primary cause is an infectious agent; for example, it is well known that genetic variance alters the susceptibility to both bacterial infections such as malaria [2,3] and viral diseases such as HIV [3], while autoimmune disease such as multiple sclerosis are also associated with specific genotypes [4]. In addition, an individual's susceptibility to the pathogenic effects of a range of exogenous toxins is strongly influenced by the activity of phase 1 and phase 2 breakdown enzymes which vary due to a number of genetic polymorphisms [5]. Even conditions which are strongly influenced by sociological forces such as eating disorders, including anorexia and obesity, are thought to have a genetic component to their pathology [6].

2. Why it is important to find functional SNPs

The number of protein encoding genes in the human genome has been found to be unexpectedly low; in the region of 20–25,000, while complexity and variation in mechanisms

* Tel.: +44 02920 744840; fax: +44 02920 744005.
E-mail address: buckland@cf.ac.uk.

0925-4439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbadis.2005.10.004

of the control of gene expression has been shown to be large. This has led to the hypothesis that variations in the expression of genes, rather than the expression of different genes, underlie the complexity and variation of human phenotypes; thus regulatory polymorphisms may be the main source of human variation [1,7,8].

In recent years, the focus of molecular genetic research has moved from rare diseases with simple patterns of inheritance to the common diseases of multi-factorial origins in which the genetic component consists of multiple genes, each of which individually makes a small to moderate contribution to disease risk. These so-called 'complex' diseases are of much greater importance to human health than the simple genetic disorders in that together they are responsible for most human morbidity and mortality (mental illness, cardiovascular disease, asthma, diabetes, dementia etc) and are probably involved in a variety of important animal phenotypes also [9,10].

Understanding the fundamental genetic mechanisms of such diseases is therefore a priority in medical research [11,12]. Central to translating the genetic findings to the clinic is identifying the pathogenic alleles. Only by identifying the pathogenic alleles can we make accurate predictions about the nature of their pathogenic function and most importantly generate accurate models of the disease in vitro and in vivo which are key steps in translating the genetic findings into new therapeutic interventions. Moreover, identifying the susceptibility allele maximises the translational power of genetics for diagnostics because without it, we cannot determining the true magnitude of the genetic effect. The susceptibility SNP itself provides optimal power for gene based risk prediction in different populations and for examination of specific phenotypic correlations. The use of the susceptibility allele is also most potent for downstream studies of gene–gene and gene–environment interactions. Understanding the latter is likely in particular to be of immense clinical value, the environment being generally more modifiable than the genome. In relation to gene regulation, the susceptibility SNP is the rSNP, which has a functional effect on expression and all rSNPs are potentially susceptibility SNPs for phenotypes, including diseases.

Although mapping disease genes for complex diseases is difficult, an increasing number of susceptibility genes have now been identified as a result of the availability of full genome sequence, dense marker maps and high throughput genotyping platforms. However, in most cases, the true susceptibility variant or variants remain unknown and extremely difficult to identify. There are several reasons for this including weak correlations between variant and phenotype as a result of weak genetic effects, allelic and locus heterogeneity, and difficulties in phenotype definition [13]. However, one of the most important issues arises from the existence of linkage disequilibrium (LD) between DNA variants. Extensive LD makes the task of gene identification easier, but it also means that it is often difficult or even impossible to distinguish on statistical grounds between the susceptibility variant and those variants simply in LD with it [8,14–16]. Moreover, it is also difficult to make such a distinction based upon the likely functional properties of the SNP. Again, this contrasts with simple genetic disorders where disease mutations are often readily identifiable because of our knowledge of the rules of gene translation and where the mutation landscape is dominated by mis-sense and nonsense mutations. However, for complex diseases, it is now generally accepted that variants which exert their effects on disease susceptibility generally do so through more subtle mechanisms, of which altering gene expression is a major player [1,7,8,11,12]. These regulatory polymorphisms or rSNPs are unfortunately not easily separated from non-functional variants based upon any predictable characteristics. Moreover, since there are an estimated 11 million human SNPs with a minor allele frequency >1% [16] as well as innumerable rarer variants that might in principle cumulatively make an important contribution to disease [17], there are no experimental procedures in use, or on the horizon, which could functionally distinguish between those of functional importance and those that are not. The sheer magnitude of the task suggests that wet-experimental procedures are unlikely to solve the problem, and that instead, an iterative combination of descriptive wet laboratory data acquisition, the development subsequently of bioinformatics predictions, and wet laboratory testing of those bioinformatics prediction will be required. The successful development of such a set of bioinformatics tools is likely to prove crucial in identifying susceptibility variants [11,12] by positional cloning either after linkage or after whole genome approaches to genetic association, the two approaches that are most powerful for those diseases where our knowledge of pathophysiology is weak and which stand most to gain from non-hypothesis based genetic approaches.

3. What variations in genes are involved?

Genetic sequence variants which involve the substitution or deletion of a single or small number of bases are frequently referred to as simple nucleotide polymorphisms, or SNPs. It is convenient to use this term regardless of the frequency with which the SNP occurs in any population or the number of bases involved.

In principal, SNPs in protein encoding genes can influence a phenotype in two ways, either by changing the quality or quantity of the encoded protein. Those changes are transmitted from the gene to the protein via mRNA; the former is represented by changes to the sequence of the encoding mRNA whilst the latter is represented by changes to the abundance of the encoding mRNA, or in the rate of translation of the mRNA into protein. Changes to the activity, processing, trafficking, etc. of the protein are controlled by other proteins, which in turn are regulated as above.

Non-synonymous SNPs are those which result in the change of a codon which gives rise to a different amino acid; however, SNPs which alter splicing can also lead to changes in protein sequence. Splicing is controlled by the splice acceptor, donor and branch site but also by exonic splicing enhancers, which are involved in the recognition of exons by the splicing protein. Changes to any of these sequences can give rise to altered

splicing and significant changes to the function of the encoded protein; up to 25% of alternative transcripts have premature stop codons leading to nonsense mediated decay [18].

Alternative splicing can occur, giving rise to alternative first exons and therefore alternative promoters. Different promoters may be used to confer tissue, developmental or state specificity as well as giving rise to different coding sequences [18]. Clearly, any SNP in an alternative promoter will only exert its effect when that promoter is used, and as such may exert a tissue, developmental or state effect. For example, a SNP in one promoter of the prolactin gene causes changes to gene expression in lymphocytes but not the pituitary, and may be associated with systemic lupus erythematosus [19].

The regulation of gene expression is complex, as all stages of the process are under some form of control. The principal control stages are at the levels of transcription, mRNA stability and translation. Any method which is based on measuring the relative abundance of mRNA which encodes specific proteins will detect changes to transcription and mRNA stability, but not translation. The control of translation is primarily modulated by sequences within the 3′ and 5′ untranslated regions (3′UTR and 5′UTR). This control has been reviewed previously [20] and the effect of sequence variants has been described [21,22] and illnesses such as myotonic dystrophy and α-thalassemia can be caused by such mutations [22].

The stability, or turnover, of mRNA is a tightly regulated process dependent on specific *cis*-acting sequences and *trans*-acting factors [23]. The minimal recognition sequence for RNA destabilising enzymes is UUAUUUAUU which appears in the 3′UTR of many genes [24]. However, many functional recognition elements do not contain this sequence or variants of it, and alone it is not sufficient for binding of the proteins involved in RNA degradation; several copies of the sequence element are required and proteins recognise and bind to a combination of primary and secondary structure elements (stem–loops) formed in the mRNA [23].

Several diseases are known to be caused by deletion of part of the recognition element [22] or by alternative splicing (e.g., [25,26] and other complex diseases may be influenced by SNPs in these elements; for example two SNPs in the 3′UTR of *CRTH2* are associated with asthma and the associated haplotype has a higher level of mRNA expression caused by an increase in mRNA stability [27]. The regulation of a number of genes is known to be affected by 3′ UTR SNPs (for example: [5,28–30].

However, there are a limited number of SNPs known to exert an effect on mRNA stability or translation. In comparison there are well over a hundred sequence variants known to affect transcription in vitro [31,32] and the presence of many more has be detected using in vivo methods [33–38]. Evidence suggests that there are many thousands more waiting to be discovered.

4. Allelic expression analysis

Factors which affect gene expression can be categorised as either *trans* or *cis*-acting phenomena. *Cis*-acting effects are those which act only on a single gene on the same chromosome, for example a SNP in a regulatory element (rSNP) or parental imprinting of a gene. *Trans*-acting influences are all other effects. Examples are: changes to one gene (regulatory or amino acid sequence) which in turn influences the expression of a second gene; and any effect resulting from a non-genetic factor such as drugs or hormones.

From the perspective of any given gene, *cis* effects are copy specific while *trans* effects influence both parental copies equally (unless the *trans*-effect is overridden by a *cis*-acting event such a polymorphism that renders one copy more or less sensitive to the *trans*-acting influence, that is, a gene–environment interaction). In practice, this makes identifying the presence of rSNPs or *cis*-effects harder than identifying *trans*-effects.

For example, the use of microarrays allows a comparison of the relative abundance of the mRNAs which encode the majority of genes in the genome, variations of which are primarily due to differences in the rate of transcription of these genes between individuals. Comparisons between groups of individuals with an illness and a control group typically identifies hundreds of differentially regulated genes. However, these differences are probably mainly due to environmental or state effects such as illness, nutritional state, age, time of day of sampling, and so on; in other words *trans*-effects. They do not reflect genetic sequence differences (*cis*-effects) in the control of those genes, or indeed of any other “upstream” genes.

A report by Morley et al. illustrates several of the points above [8]. The variation in expression of ~8500 genes in CEPH lymphoblast cell lines was studied using microarrays. Using the known genotypes of these samples Morley et al. (2005) were able to carry out in silico linkage analysis. They found that the basal expression of up to 1000 genes appeared to be controlled by genetic effects, and of 142 studied in more detail, 20% appeared to have a *cis*-acting transcriptional regulator. However, they were not able to indict any specific SNPs as being rSNPs.

Recently, we [33] and others [34–38] have developed indirect methods for detecting the presence of even distal *cis*-acting sequence variants that affect the expression of an assayed gene independently of *trans* effects. By measuring the relative expression of the paternal and maternal copies of autosomal genes we can selectively detect *cis*-effects, heterozygosity for which is indicated by deviation of the ratio of expression of parental copies from unity, while controlling precisely for *trans* effects.

The detection of allele-specific gene expression in a single individual relies on the ability to distinguish the gene product of one parental chromosome from that of the other, and then to quantitate the relative amounts of each gene product that is produced. Three studies have been carried out using the method of single nucleotide primer extension [39]. A transcribed polymorphism is used as a marker to distinguish between the mRNA products of the parental chromosomes. The relative abundance of each allele from a heterozygous individual is then quantitated using RT-PCR and primer extension with radio labelled [39] or fluorescent nucleotides

[33,34,38]. Both gene copies come from the same tissue samples and have been subject to the same environmental influences including genetic *trans*-acting factors and experimental insults including mRNA degradation. In the absence of either *cis*-acting sequence variation or epigenetic effects affecting expression of the target mRNA, each chromosome should be equally expressed regardless of the absolute level of gene expression. The ratio of the abundance of each allele is therefore expected to be ~1. In samples that are heterozygous for a *cis*-acting regulatory variant or epigenetic modification, mRNA originating from one chromosome will be expressed at a higher level than that from its sister chromosome and this is detected by changes in the ratio of abundance of each mRNA allele.

In the first reported survey of allelic discrimination, this method [39] was used to study 13 genes in 96 CEPH lymphoblast samples and variable allelic ratios were found in 6 genes [34]. The variations were found in 18–39% of individuals and the magnitude of variation ranged from 1.3- to 4.3-fold, although only 1 gene gave variations above a factor of 1.9. Of importance, the results from three families were consistent with inheritance of the variations.

A second survey used a similar method to that above to study an additional 15 genes [33]. Of importance, the mRNA was extracted from 50 post mortem human brain cortex samples, eliminating the possibility of artefacts caused by the immortalization of lymphocytes. Seven genes displayed allelic expression differences of over 1.2-fold, the largest being 1.7. For the variably expressed genes, between 5% and 66% of individuals showed differences.

A third survey again used a similar method to that above to study 129 genes in lymphoblast mRNA [38]. Of those, 23 genes showed relative allelic expression of 1.7-fold or greater (the chosen cut off point), the largest difference, not including imprinted genes, being 9-fold. For the variably expressed genes, between 11% and 100% of individuals showed differences. At least one of these genes was also shown to have allele-specific expression in adipose tissue.

The similarity in these three studies of the magnitude of changes and the proportion of both the genes and individuals showing variation suggests that the source of the tissue is not critical, although clearly any state effects related to illness and drug treatment could only be detected in primary tissue samples.

A higher through-put method, allele-specific micro-arrays has been used to measure the levels of mRNAs [37]. 602 genes were studied and 326 (54%) showed a greater than 2-fold difference, while 170 (28%) showed a greater than 4-fold difference. However, this method only allowed the identification of 2-fold or greater differences in allelic expression. The work was carried out using liver and kidney from a relatively small number of human fetuses. For each gene, no more than 5 heterozygotes were studied and for over 220 (68%) of the genes only 1 or 2 heterozygotes were available. Thus, they had very low power to detect allelic differences and the rate of differences found is therefore far higher than that found by others above [33,34,38] and it is not clear whether the higher frequency and magnitude of variation stem from the tissue differences or experimental limitations.

An alternative approach to measure allele-specific transcription has been developed which measures differential initiation of transcription from genomic DNA using a SNP in the promoter region as a marker [36]. An antibody specific to Ser5 phosphorylated DNA pol II is added to the samples and crosslinked to the DNA by formaldehyde treatment. The chromatin is fragmented and the resulting protein bound chromatin is immunoprecipitated (ChIP). The relative amount of each immunoprecipitated allele is measured using PCR and primer extension. Despite the power of this method, only a small number of genes have been studied this way [36,40–42].

While the above experiments do not allow an exact determination of the proportion of genes which contain a regulatory polymorphism (rSNP) in any population, what can be concluded is that in a small population of less than 100 individuals at least 50% of the genes have an rSNP which alters transcription by 20% or more. This is almost certainly an underestimate as a variety of tissue-specific effects would not be seen in a limited number of tissues and many state-specific effects may not be found in such a small population. In addition, as more than half of all sequence variants are rare (minor allele frequency; MAF <1%) then in a small population as above only about half of all SNPs present in a large population would be found. These all suggest that given a large enough population that for the vast majority of genes one or more individuals will carry an rSNP.

However, these approaches, based upon allele-specific quantitation of gene expression, have major limitations. Of these, the most important is that they do not identify the specific regulatory variant or variants responsible for altered allelic expression, although as we have shown, in some instances functional haplotypes can be identified [43–45]. However, even where this is possible, it is still necessary to use an in vitro method such as a reporter gene assay in order to implicate specific *cis*-acting polymorphisms as causal for altered regulatory function. To achieve this, specific polymorphisms in specific regulatory elements have to be identified and isolated.

A recent review describes various in vitro methods which can be used to identify or validate regulatory SNPs [46]. One of the methods which has not been widely used for this purpose but which has great potential is the electrophoretic mobility shift assay (EMSA). Many proteins involved in regulation of gene transcription bind directly to specific DNA sequences in order to exert their effect. Historically, many such DNA sequences have been identified using EMSA. Double stranded DNA of the sequence to be studied (which may be synthetic oligonucleotides or short-cloned strands of DNA) is tagged with a radioactive or fluorescent label and incubated with cellular extracts to allow any DNA binding protein to bind if its recognition sequence is present. The mixture is analysed on a polyacrylamide gel and two bands should be present, one representing the unbound DNA and a second band at higher molecular weight which represents the DNA bound to a protein. If two alleles of the same DNA segment are analysed

separately, the relative ability of each sequence to bind to the protein can be assessed. A recent study by Mottagui-Tabar et al. [47] has shown that this method could be used as the first step in a procedure to find regulatory SNPs. They studied 10 SNPS which were located within putative transcription factor binding sites (TFBS) and found that 7 of them did indeed give differential protein binding. The effect of 4 of these 7 SNPs was subsequently confirmed using a reporter gene assay.

Linnell et al. [48] have published a high throughput method which gives similar information to a quantitative EMSA. By using oligonucleotides on chips they cannot only quantitatively determine the relative affinities of different alleles but by comparing a large number of different sequences, they can go some way to defining the influence of each sequence position on the affinity of the TFBS for the respective transcription factor (TF).

5. Reporter gene assays

The core promoter and proximal promoter regions contain the elements that control the initiation of transcription and are, therefore, a priori regions that plausibly harbour functionally relevant polymorphisms with major effects on gene expression. Also, at present, of the possible regulatory elements affecting the transcription of a gene, only the core and proximal promoter elements have highly specific predictable spatial relationships with their respective genes, a relationship that we [49] and others [50] have shown allows promoter sequences to be identifiable in most cases, despite imperfections in gene annotation, from public databases. Several in silico promoter prediction methods have been developed in the past but with limited predictive performance due to the lack of any consensus sequences or structural features which are both promoter specific and found in all promoters [51]. More recently a number of advances in bio-informatics have allowed more accurate identification or promoters [52]: http://www.rulai.cshl.edu/CSHLmpd2/; [53]: http://www.mgs2.bionet.nsc.ru/argo/; [54] The newer approaches are based upon the complete human genome sequence, comparison of with that of other species, full length mRNA sequences and also machine learning algorithms. This makes promoters the only regulatory element that can at present be studied in a high-throughput manner and also the obvious first place to look for functional polymorphisms.

The methodology most widely used to study gene promoters has been the reporter gene assay. In brief, this involves cloning different alleles of the regulatory sequence being studied (usually the promoter region) into a plasmid which contains the coding sequence for a reporter protein which is easily quantifiable, such as luciferase, in such a way that the relative activity of the regulatory sequence can be determined by the amount of the reporter protein which is produced following transfection of the plasmid into cultured cells [55].

Other sequences, which contain regulatory elements can be studied using reporter gene assays, if they can be identified, but the size of the region which may contain rSNPs for a gene, may be 100s of kb long. The size of sequence which can be studied is limited not only by the cloning technology, but by the number of SNPs likely to be found in the sequence. If there is more than one bp difference between two sequences which are to be compared, any difference in transcriptional activity cannot be assigned to a specific locus unless all possible haplotype variants have been analysed, and possible multiple effects and epistasis have been taken into account. A second problem related to size are Taq DNA polymerase induced base changes, the frequency of which increases with size of the amplimer and very long sequences are hard to copy with 100% fidelity.

Promoter–reporter gene assays have been used since the early 1980s and a large number of promoter region SNPs have been analysed [31]. Rockman and Wray [31] list approximately 100 SNPs which have given positive results in reporter gene assays [31]. However, as most SNPs to be tested were pre-selected on the basis of having prior probability or functionality, a true assessment of the proportion of promoter SNPs which are functional could not be carried out. In addition, differences in experimental technique and thresholds used make the data less amenable to meta-analysis than would be desired.

The author and colleagues have recently completed a large scale project to identify functional promoter SNPs in a systematic manner [32]. This project had the advantages over a meta-analysis as described above:

- SNPs studies were all those found in a screening set of 32 chromosomes, with no pre-selection.
- The same techniques and procedures were used throughout, including the use of the same cell lines.
- The same threshold for functionality was employed.

Approximately 1000 genes were selected either randomly, by chromosome (Chr 21 and 22-selected as they were the first 2 chromosomes to be sequenced allowing a large proportion of the genes to have their promoters identified) or by relevance to mental illness [32,56–62]. Using a screening set of 16 individuals of diverse ethnic origin, to find SNPs in the first 500 bp upstream of the reported start of transcription, half of the genes analysed (341) had 1 or more SNPs. We were able to clone and functionally analyse 247 of the genes using two different cell lines. We found that 54 genes had promoter haplotypes which differed in their ability to promote transcription by 50% or more. This threshold was chosen as it is equivalent to the presence of a third copy of the gene.

We demonstrated that 8.4% of naturally occurring sequence variants in the first 500 bp 5′ to the start of transcription resulted in changes in transcriptional activity in vitro by over 1.5-fold [32,56]. From the perspective of the gene rather than the polymorphism, even in our small sample set of 32 chromosomes, functional variants were detected in approximately 10% of the genes studied. Given our study focussed only on one regulatory element, the promoter region, that we only had 30% power to detect any given allele with a minor frequency of 1%, and that our definition of a functional

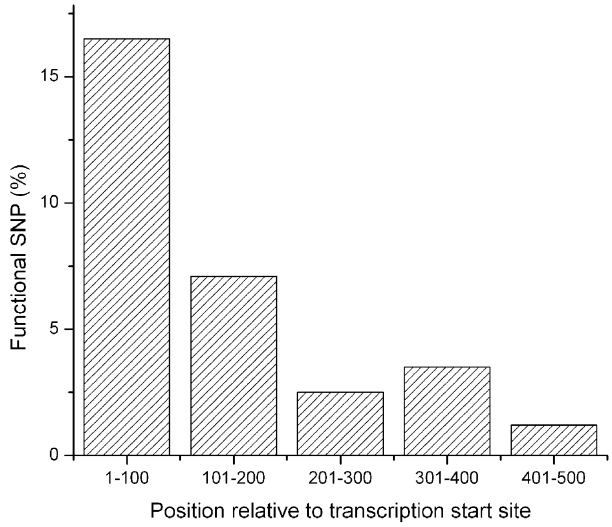

Fig. 1. Percentage of all sequence variants that are functional in relation to the distance from the transcription start site. Sequence variants found in gene promoters with more than one matching locus in the human genome were excluded. Reprinted with permission from [32].

difference is very conservative [56], clearly, the true number of genes whose expression is influenced by promoter region polymorphisms is much higher than this.

Forty SNPs were identified as having functionality independently of other SNPs, 20 of these lay within 100 bp of the TSS while another 10 lay within 100–200 bp of the TSS. The non-functional SNPs were distributed evenly showing that the position of functional SNPs is strongly biased towards the core and proximal promoter as shown in Fig. 1.

This may be due to the fact that the RNA polymerase II binding site and other TF binding sites are present in this region of the promoter. However, an analysis of the sequences within which the functional SNPs lie shows that the majority do not appear to disrupt a known TF binding site. This is unexpected as the core and proximal promoter regions are probably the best studied regulatory regions of human genes and the existence of a large number of unknown consensus sequences which lie within this region seems unlikely. Indeed, this has been demonstrated for sequence of eight base or less. In a large survey of 13,000 human promoters, Fitzgerald et al. (2004) determined the distribution of each of the 65,536 possible eight base DNA sequences which might act as TFBSs [63]. They found statistically significant clustering of seven known TFBS consensus sequences and a novel sequence they named as Clus1. Each of these was mainly found within 100 bp of the start of transcription. This suggests that if unknown TFBSs in the core promoter region are waiting to be discovered there must be a large number of relatively rare consensus sequences, rather than a few common ones. However, a more recent survey which studied sequences of 6 to 18 bp by comparison with other species, found 105 new motifs in human promoters, 85 of which were 9 or more bp long [64] and therefore would not have been found by Fitgerald et al. [63]. It is not known if any TFs bind to these sequences or if they have other roles in the regulation of transcription.

An additional finding was that 29 of the gene promoters which contained a functional SNP were active in both cell lines; 18 of these SNPs were functional in both cell lines while 11 were functional in only one cell line. Cell line specificity is most likely due to the presence of a TFBS for a tissue-specific TF. Interestingly, when SNPS with cell line specificity were compared with those laying with TFBSs, a strong correlation was observed (Table 1). This result therefore, is consistent with the hypothesis that those functional sequence variants not in a known TFBS may influence transcription by a mechanism which is independent of the presence of transcription factors, possibly related to some basic properties of the DNA itself.

Overall, our research and that of others has shown genes are commonly influenced by *cis*-acting polymorphisms, and that as we predicted, the promoter is a rich source of functional variants.

However, there are a number of important questions which stem from that work.

5.1. What is the relationship between in vitro results and the situation in vivo?

Allelic expression assays are powerful in detecting the presence of *cis*-acting effects. However, they have limited utility in locating or identifying the source of the effect. For this, the method of choice is the reporter gene assay. In a small number of cases changes found in reporter gene assays have been causally associated with a functional effect [1], but in the overwhelming majority of cases there is no evidence that these in vitro effects have any bearing on in vivo expression levels. Reporter gene assays use only a limited amount of the genomic DNA sequence associated with a gene and any effects stemming from the sequence analysed may be abolished or modulated by the effects of other sequence elsewhere in the genome. In addition, the structure of the chromatin and the effects of nucleosomes are not taken into account.

While it is not unreasonable to assume that reporter gene assays are, in fact, relevant to in vivo expression, these experiments are of such fundamental importance in the study of gene expression that an investigation into the reliability and validity of these assays is overdue. The obvious way to do this would be by comparing data between reporter gene and allelic expression assays. This is not as straight forward as it seems. Given an mRNA sample from an individual who is heterozygous for a promoter region SNP which is functional in an in

Table 1
Correlation between tissue specificity and context of functional SNPs

Functionality	In TF binding site	Not in TF binding site
Both cell lines[a]	4	14
Cell line specific	6	5

Correlation between SNPs which exert a tissue-specific effect and whether or not they fall within a known transcription factor binding site consensus sequence (nucleotide sequence context). Eleven SNPs in promoters were inactive in one cell line and are not included. Reprinted with permission from [32].

[a] Includes SNPs which lead to a statistically significant activity between alleles of 1.3 fold or greater, $P=0.11$, 2 tail (Fishers exact test).

vitro assay, the in vivo allelic expression analysis of the two gene copies compares at the same time all sequence differences in the gene region. Therefore, any difference in expression may be related to any one of these. If a number of heterozygous and homozygous samples are similarly analysed, in principal, if all the heterozygotes give a difference in the same direction while none of the homozygotes do, then this suggests that the SNP in question is functional. However, it may be in linkage disequilibrium with another functional SNP and in addition, there may well be other functional SNPs which would give apparent "false positives" and "false negatives", the latter occurring when the effect of the target SNP is countered by another functional SNP.

5.2. Does differential allelic expression lead to phenotypically relevant changes?

In vivo analyses described above suggest that at least 50% of genes show differential allelic expression. However, although one haplotype of a gene may be expressed at a lower level *relative* to another in individuals who are heterozygous for that haplotype (confirming a *cis*-effect of that haplotype), *trans*-acting homeostatic mechanisms might compensate for this, resulting in unaltered total mRNA abundance. If this is the case for any rSNP, it may have much less phenotypic impact that might be presumed. The report by Morley et al. [8] discussed above shows that the expression of many mRNAs is influenced by genotype, but give no measure of the proportion of rSNPs which do change total mRNA levels, nor do they identify the rSNPs.

Similarly, Hirota et al. showed a correlation between the total *CYP3A4* mRNA level and allelic expression ratio, but did not identify the rSNP(s) involved [65]. However, Liu et al. showed by chromatin immunoprecipitation [36] that the *GSTM3* gene has a promoter rSNP (−63A/C) which changes mRNA levels, as shown by quantitative PCR [42]. It remains to be seen what proportion of the rSNPs do actually exert a phenotypic effect, and what the relative size of the two effects is.

5.3. What are the mechanisms by which promoter sequence variants exert their effects on transcription

Several bio-informatic approaches to determining whether or not a known SNP is functional have been reported, but all depend on the SNPs disrupting or creating a TFBS [66–68]. Analysis of the sequence variants that influenced expression in the reporter gene assay from the Cardiff promoter project [32] revealed that most (~70%) were not associated with known consensus sequence motifs for TFBSs. This suggests that either there are many as yet unknown consensus sequences relevant to TF binding or that the majority of functional sequence variants affect gene transcription in other ways. At a fundamental level, the potential importance of determining the answer to this question is equivalent to that of knowing the role of codons in understanding the role of mis-sense and nonsense mutations. Such an understanding is likely to be decisive if we are to develop methods for predicting the functional properties of the vast majority of promoter regions SNPs, and possibly SNPs in other regions adjacent to genes, whose function is likely to remain un-assessed experimentally given the laborious nature of doing so, at least with current technology.

6. The human core promoter

The data presented in Buckland et al. (2005) [32] and Fig. 1, as described above, suggests that the core and proximal promoter regions are particularly sensitive to changes of a single base pair. However, it is not clear why this should be so; is it because there is a higher concentration of TFBSs or are there other factors playing a role?

Text book descriptions of human core promoters, the region which binds to the DNA polymerase II complex, usually show a number of consensus sequences with the implication that most if not all promoters contain these elements, especially the TATA box. TATA-containing promoters were first discovered in bacterial genomes and the TATA box was thought to be the universal promoter element. Human TATA-less promoters were discovered later and their percentage in the total number of studied promoters has risen steadily since: from 22% [69] to 36% [70] to 68% [71] and most recently to 78% [72]. The TATA box is therefore present in only a fifth of human promoters and in the others, RNA Pol II must recognise some other factor(s) in order to correctly initiate the start of transcription. Several other core promoter elements are also known which fulfil this role: the TFIIB recognition element (BRE), the downstream promoter element (DPE) and the initiator (Inr). Each of these is only found in a proportion of human promoters, the most common element being the Inr which is present in approximately 50% of promoters [72] but any combination of one, two or three of the elements may be present, with or without a TATA box. Each of these elements acts, either alone or with the others to recruit TFIID to the polymerase initiation complex, by binding to another transcription factor which in turns binds TFIID, with the exception of Inr which binds TFIID directly. Where the above elements are found, the spacing between them can have a marked effect on transcription [72]. It is also possible that sequences in the neighbourhood of these specific elements are also be involved in the TFIID binding process. However, approximately 25% of promoters have none of the elements above and no other TFIID binding sequences are known. This suggests that other factors are involved in TFIID binding. One possibility is that other TFs bind to their recognition elements and direct TFIID to the correct site. Therefore, disruption of these TFBSs would affect transcription. Another possibility is that the second order properties of the promoter sequence can also play a role in transcription initiation.

7. Transcription factor binding sites (TFBS)

At present, determining whether or not a base lies within the DNA binding site for a protein transcription factor is fraught

with difficulty [73]. Firstly, it is possible that many TFBS consensus sequences are not yet known. However, even if we assume all are known, determination of the presence of a TF binding site and the effect a base change has on its affinity for the TF is still problematic. TFBSs are not simply defined by a specific sequence of nucleotides but rather by a set of such sequences where the bases at several of the positions are degenerate. In practice, there are very few human promoter sequences that exactly match the consensus sequences. To allow for this, sequences are interrogated for the presence of putative TFBS using a matrix allowing for possible degeneracy at all sites [63,73]. However, many weighting matrices are derived from a limited number of known binding sites and the effect of variation and the exact weight that should be accorded for each of the nucleotides independently when predicting a TFBS consensus sequence is not known for most such sites [73,74]. Moreover, the affinity of the TF for its binding site is not simply dictated by additive effects of changes at each site, with nucleotide changes being interdependent in their effects [74]. This means that in some cases at least, the effect of changes to TF sites can only be determined experimentally for each and every possible oligonucleotide sequence [74]. While work to do this is being carried out [75], it will be some time before the effect of all TFBS SNPs is known.

8. DNA structure

There have been few investigations into the relative compositions of strong and weak eukaryotic promoters. The presence of TFBSs clearly play a role, but several studies, both experimental and computational, have shown that promoter regions possess a number of sequence-dependent characteristics which make them distinct from the rest of the genome; stability, curvature and flexibility [76]. Curvature refers to the shape of the DNA sequence and how it is bent. Sequence flexibility refers to the ease with which the DNA can bend to allow interactions of proteins bound to the DNA in different places, or to avoid steric hindrance. Stability refers to the ease with which the DNA can become single stranded to allow transcription. While the majority of studies on the effects of DNA structure on transcription have been carried out using non-human DNA, there is no reason to believe that similar mechanism do not apply to human DNA and this is supported by some evidence [76].

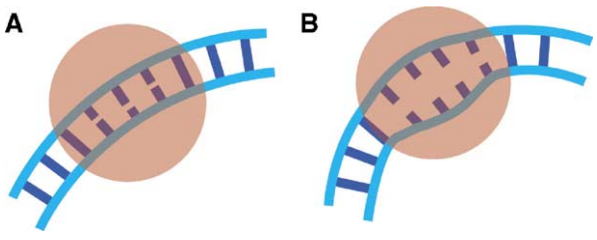

Fig. 2. DNA melting following bending. When a transcription factor such as the TATA binding protein binds to a DNA consensus sequence, it can cause the DNA to bend and the strands to separate, as in (B) allowing transcription to commence. If the formation of the protein–DNA complex does not result in sufficient bending, strand separation does not occur and transcription is inhibited as in (A).

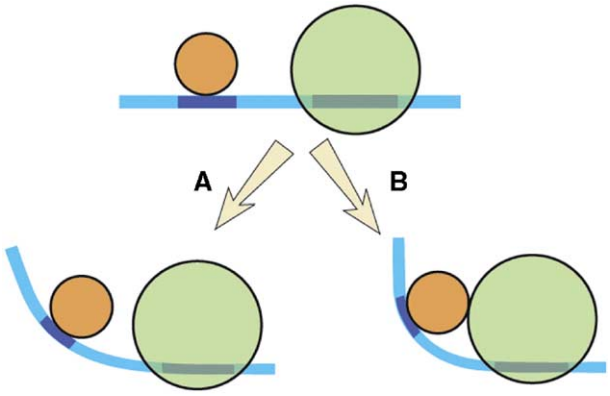

Fig. 3. Flexible DNA allows transcription factor interaction. To allow two or more protein–DNA complexes, which are not adjacent, to interact, the DNA needs to bend around. The more flexible the DNA the greater the ease with which the interaction can occur.

Pedersen et al. studied the prokaryotic genome and found that in practically all the investigated eubacterial and archaeal genomes, there is a trend for promoter DNA to be more curved, less flexible, and less stable than DNA in coding regions and in intergenic DNA without promoters [77]. This trend is present regardless of the absolute levels of the structural parameters, and they suggest that this may be related to the requirement for helix unwinding during initiation of transcription.

8.1. DNA curvature

Curvature of the DNA duplex is central to how many TFs exert their effects on gene expression. Depending on its sequence, curvature can be an inherent property of a DNA molecule or it can be induced by external factors, such as protein binding [78]. It is known that curved DNA is often located near transcriptional control regions [79,80].

The affinity of a transcription factor for its recognition sequence is influenced by the average curvature of the unbound DNA and there are different mechanisms leading to this effect. One effect of curved DNA has been highlighted [80]; intrinsically curved DNA is often found upstream of the TATA box and it has been shown that this promotes an association with nucleosomes which bends DNA towards the major grove in the TATA region, exposing the minor grove and this increases the rate of transcription. The curved DNA need not itself be part of a TFBS to have this effect [80]. Kim et al. [81] have shown that the insertion of one or more copies of an intrinsically curved DNA sequence (A_6CGTG) into a minimal promoter upstream of a TATA box increases transcription by 100-fold [81]. Change to this sequence which abolished DNA bending reduced transcription by 70%. They also showed that a sequence specific, rather than bending specific protein binds to this sequence and promotes transcription. We have found a similar sequence to that used by Kim et al. [81] which occurs naturally in the promoter of

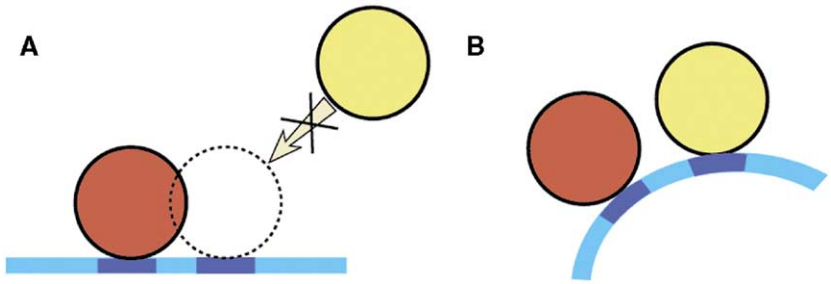

Fig. 4. DBA bending alleviates steric hindrance. Where two transcription factors are required to bind to their respective binding sites and the latter are close together, they may only be able to bind if the DNA is bent to give more room. (A) Steric hindrance stops both proteins binding at the same time. (B) The curved DNA allows more space for both proteins to bind at the same time.

the *PEDF* gene which is important in the development of the retina. Interestingly, a promoter reporter gene clone with the sequence A_6GCTC, has twice the transcriptional activity of a different allele having the sequence A_6GATC [32]. Although speculative, the AT dinucleotide is less flexible than the CT dinucleotide [82] and this may influence the curvature of the sequence.

A second effect of curved DNA stems from the fact that binding of many TFs to DNA results in the bending of the DNA. Unbound DNA with a curvature similar to that of the DNA–TF protein complex has a lower free energy of DNA–TF protein association (and therefore higher affinity for the TF) than unbound DNA whose curvature is relatively dissimilar, and which requires more distortion after TF binding [83].

Experimental evidence suggest that following binding of the RNA pol II to the promoter, the DNA strands separate and wrap around the polymerase causing a sharp bend in the DNA [84]. Also, the TATA binding protein (TBP) binds the minor groove of the TATA recognition sequence and distorts the DNA 80° towards the major groove. DNA with a natural curvature towards the major groove binds TBP with 100-fold greater affinity than linear DNA and with 300-fold greater affinity than DNA bent towards the minor groove [85]. An illustration of the probable mechanism of this effect is shown in Fig. 2. Many other transcription factors also facilitate the adoption of curved conformations by DNA molecules [86,87]. Examples include the cAMP receptor protein (CAP), purine repressor (PurR), integration host factor (IHF) and the TATA-binding protein (TBP) [78].

Another example of a TF that exerts its effect through bending DNA is yeast MCM1. Mcm1 is a member of the MADS-box family of TFs which are found in eukaryotes, including yeast and mammals [88] and is involved in regulating a range of genes. Factors which influence the degree of DNA bending have been shown to be critical for the activation of transcription following TF binding to the consensus TFBS. A mutation in the MCM1 TF itself changes its ability to bend DNA from 95° to 82° and this lowers transcription by >10-fold [89,90]. Mutations in single bps of the TFBS for MCM1 that are known to be involved in bending, also decreases transcription by 4-fold [91,92]. However, although these substitutions have a large effect on DNA bending and transcriptional activation by Mcm1, they have a relatively small effect on the DNA-binding affinity of the protein.

Perhaps with respect to a human phenotype, one of the most dramatic consequences of curvature and bending comes from a mutation to the human male sex-determining factor SRY, a mammalian TF. The High Mobility Group (HMG)-box domain is approximately 80 residues in length and defines a superfamily of architectural factors that play a central role in mammalian gene regulation and organogenesis [93]. The HMG-box domain of SRY binds to specific DNA sequences in the minor groove, resulting in substantial DNA bending. A number of mutations in the SRY gene which give rise to amino acid changes in the DNA binding domain result in 46X,Y sex reversal. One particular mutation has been shown to change the DNA bend angle by 13°.

8.2. DNA flexibility

Flexibility of a DNA molecule can be defined as the ease with which the molecule can be made to curve in any direction, and this is dependent on its sequence [94]. DNA is an inherently stiff molecule; however, some DNA sequences, while being intrinsically straight, can readily undergo distortion. The flexibility of DNA has been shown to be important in the formation of many protein–DNA complexes [94] and increasing the flexibility may allow greater binding of a TF or RNA pol II [95]. Examples of this include TATA Binding Protein [85,96] and Catabolite Activator Protein [97].

There are a number of ways in which flexibility of DNA influences the way TF interact with DNA and modulate transcription: DNA does not easily loop around to allow upstream enhancer elements to become adjacent to the DNA

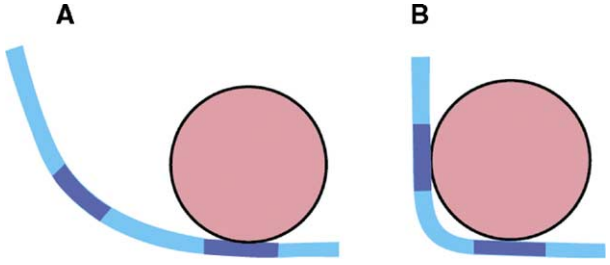

Fig. 5. To allow one protein to bind to two stretches of DNA curved or flexible sequences between the binding sites may be required as shown in B.

Pol II site as is often depicted in text books [98]; rather stretches of flexible DNA facilitate this as shown in Fig. 3. DNA sequences that are not intrinsically at the optimal curvature for the binding of a TF but that can be easily distorted by TF binding to it as a result of greater intrinsic flexibility have lower free energy of DNA–TF association than more rigid structures [83]. This kind of "indirect recognition" may be more important in TF binding when the complex requires large distortions in the DNA structure [83]. Flexibility of DNA is important when two TFs bind to TFBS which are close together, which is more likely to occur in the proximal promoter where many TFBS are found. If the DNA is rigid, both TFs may not be able to bind at the same time due to steric hindrance (both proteins are trying to occupy the same physical space at the same time). If the DNA is flexible, it can bend to allow both TFs to bind at the same time [87] as depicted in Fig. 4. Alternatively, DNA may be required to bend to allow a TF to interact with two binding sites as shown in Fig. 5.

8.3. DNA stability

DNA stability or its resistance to becoming single stranded depends primarily on the sum of the interactions between the constituent mono and dinucleotides. (For an oligonucleotide this is referred to as the melt temperature.) Low stability aids the melting of DNA by DNA polymerase and thus aids transcription. Eukaryotic core and proximal promoter regions are less stable than coding regions and often show a peak of minimum stability between the −25 and −35 region, probably due to the presence of the TATA box in some promoters [76]. This can easily be determined bio-informatically [99]. A simple analysis of the functional SNPs found in the Cardiff Promoter Project found no resulting large changes in DNA stability and approximately 50% of the SNPs increased the stability (data not shown). Changes to DNA stability does not therefore appear to be a major mechanism by which promoter SNPs exert their effect.

9. Conclusions

The number of functional sequence variants in the human genome, which are relevant to a phenotype, is unknown, but may be many thousands. Identifying which of the many millions of variants are functional is important for health and research, but clearly each variant cannot be tested experimentally. What is required is a bio-informatic method to assess the likelihood of functionality based upon extensive experimental data. However, at this time, the mechanism of action of few functional SNPs in known and even those which appear to lie within a transcription factor binding site may not in fact exert their effect by its disruption. However, although structural factors such as DNA curvature and flexibility are known to have a critical effect on promoter function, little if any research has been carried out to show if a single nucleotide polymorphism can change DNA structure such as to have a functional effect. Nevertheless, in the absence of competing explanations, my hypothesis, is that sequence variants which lie outside of the consensus TF binding sites may exert their effects by either altering the bending of the DNA towards (or away from) its optimum configuration, or by altering the flexibility of the DNA and changing its ability to bend in the required way. The differences in level of transcription that have been found to result from *cis*-acting effects are mainly small. Thus, they may be caused by relatively small changes to the conformation of DNA.

References

[1] J.C. Knight, Regulatory polymorphisms underlying complex disease traits, J. Mol. Med. 83 (2005) 97–109.
[2] G. Min-Oo, P. Gros, Erythrocyte variants and the nature of their malaria protective effect, Cell. Microbiol. 7 (2005) 753–763.
[3] A.J. Frodsham, A.V.S. Hill, Genetics of infectious diseases, Hum. Mol. Genet. 13 (2004) R187–R194.
[4] D.A. Dyment, B.M. Herrera, M.Z. Cader, C.J. Willer, M.R. Lincoln, A.D. Sadovnick, N. Risch, G.C. Ebers, Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance, Hum. Mol. Genet. 14 (2005) 2019–2026.
[5] A.D. Johnson, D. Wang, W. Sadee, Polymorphisms affecting gene regulation and mRNA processing: broad implications for pharmacogenetics, Pharmacol. Ther. 106 (2005) 19–38.
[6] C.G. Fairburn, P.J. Harrison, Eating disorders, Lancet 361 (2003) 407–416.
[7] P.R. Buckland, Allele-specific gene expression in humans, Hum. Mol. Genet. 13 (2004) R255–R260.
[8] M. Morley, C.M. Molony, T.M. Weber, J.L. Devlin, K.G. Ewens, R.S. Spielman, V.G. Cheung, Genetic analysis of genome-wide variation in human gene expression, Nature 430 (2004) 743–747.
[9] S.C. Liefers, R.F. Veerkamp, M.F. te Pas, C. Delavaud, Y. Chilliard, M. Platje, T. van der Lende, Leptin promoter mutations affect leptin levels and performance traits in dairy cows, Anim. Genet. 36 (2005) 111–118.
[10] J.D. Nkrumah, C. Li, J. Yu, C. Hansen, D.H. Keisler, S.S. Moore, Polymorphisms in the bovine leptin promoter associated with serum leptin concentration, growth, feed intake, feeding behavior, and measures of carcass merit, J. Anim. Sci. 83 (2005) 20–28.
[11] N. Risch, Searching for genetic determinants in the new millennium, Nature 405 (2000) 847–856.
[12] T.J. Hudson, Wanted: regulatory SNPS, Nat. Genet. 33 (2003) 439–440.
[13] M.C. O'Donovan, N.M. Williams, M.J. Owen, Recent advances in the genetics of schizophrenia, Hum. Mol. Genet. 12 (2003) R125–R133.
[14] J.D. Rioux, M.J. Daly, M.S. Silverberg, K. Lindblad, H. Steinhart, Z. Cohen, T. Delmonte, K. Kocher, K. Miller, S. Guschwan, E.J. Kulbokas, S. O'Leary, E. Winchester, K. Dewar, T. Green, V. Stone, C. Chow, A. Cohen, D. Langelier, G. Lapointe, D. Gaudet, J. Faith, N. Branco, S.B. Bull, R.S. McLeod, A.M. Griffiths, A. Bitton, G.R. Greenberg, E.S. Lander, K.A. Siminovitch, T.J. Hudson, Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease, Nat. Genet. 29 (2001) 223–228.
[15] C. Francks, S. Paracchini, S.D. Smith, A.J. Richardson, T.S. Scerri, L.R. Cardon, A. Marlow, I.L. MacPhie, J. Walter, B.F. Pennington, S.E. Fisher, R.K. Olson, J.C. DeFries, J.F. Stein, A.P. Monaco, A 77 kilobase region of chromosome 6p22.2 is associated with dyslexia in families from the UK and USA, Am. J. Hum. Genet. 75 (2004) 1046–1058.
[16] D.A. Hinds, L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, D.R. Cox, Whole-genome patterns of common DNA variation in three human populations, Science 307 (2005) 1072–1079.
[17] J.C. Cohen, R.S. Kiss, A. Pertsemlidis, Y.L. Marcel, R. McPherson, H.H. Hobbs, Multiple rare alleles contribute to low plasma levels of HDL cholesterol, Science 305 (2004) 869–872.
[18] S. Stamm, S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T.A. Thanaraj, H. Soreq, Function of alternative splicing, Gene 344 (2005) 1–20.

[19] A. Stevens, D. Ray, A. Alansari, A. Hajeer, W. Thomson, R. Donn, W.E. Ollier, J. Worthington, J.R. Davis, Characterization of a prolactin gene polymorphism and its associations with systemic lupus erythematosus, Arthritis Rheum. 44 (2001) 2358–2366.
[20] G.S. Wilkie, K.S. Dickson, N.K. Gray, Regulation of mRNA translation by 5′- and 3′-UTR-binding factors, Trends Biochem. Sci. 28 (2003) 182–188.
[21] J.T. Mendell, H.C. Dietz, When the message goes awry: disease-producing mutations that influence mRNA content and performance, Cell 107 (2001) 411–414.
[22] B. Conne, A. Stutz, J.D. Vassalli, The 3′ untranslated region of messenger RNA: a molecular 'hotspot' for pathology? Nat. Med. 6 (2000) 637–641.
[23] A. Bevilacqua, M.C. Ceriani, S. Capaccioli, A. Nicolin, Post-transcriptional regulation of gene expression by degradation of messenger RNAs, J. Cell. Physiol. 195 (2003) 356–372.
[24] A.M. Zubiaga, J.G. Belasco, M.E. Greenberg, UUAUUUAUU is the key AU-rich sequence motif mRNA degradation, Mol. Cell. Biol. 15 (1995) 2219–2230.
[25] B. Chowdhury, C.G. Tsokos, S. Krishnan, J. Robertson, C.U. Fisher, R.G. Warke, V.G. Warke, M.P. Nambiar, G.C. Tsokos, Decreased stability and translation of T cell receptor zeta mRNA with an alternatively spliced 3′-untranslated region contribute to zeta chain down-regulation in patients with systemic lupus erythematosus, J. Biol. Chem. 280 (2005) 18959–18966.
[26] W. Farris, M.A. Leissring, M.L. Hemming, A.Y. Chang, D.J. Selkoe, Alternative splicing of human insulin-degrading enzyme yields a novel isoform with a decreased ability to degrade insulin and amyloid beta-protein, Biochemistry 44 (2005) 6513–6525.
[27] J.L. Huang, P.S. Gao, R.A. Mathias, T.C. Yao, L.C. Chen, M.L. Kuo, S.C. Hsu, B. Plunkett, A. Togias, K.C. Barnes, C. Stellato, T.H. Beaty, S.K. Huang, Sequence variants of the gene encoding chemoattractant receptor expressed on Th2 cells (CRTH2) are associated with asthma and differentially influence mRNA stability, Hum. Mol. Genet. 13 (2004) 2691–2697.
[28] C. Kealey, K.S. Brown, J.V. Woodside, I. Young, L. Murray, C.A. Boreham, H. McNulty, J.J. Strain, J. McPartlin, J.M. Scott, A.S. Whitehead, A common insertion/deletion polymorphism of the thymidylate synthase (TYMS) gene is a determinant of red blood cell folate and homocysteine concentrations, Hum. Genet. 116 (2005) 347–353.
[29] I. Puga, B. Lainez, J.M. Fernandez-Real, M. Buxade, M. Broch, J. Vendrell, E.A. Espel, Polymorphism in the 3′ untranslated region of the gene for tumor necrosis factor receptor 2 modulates reporter gene expression, Endocrinology 146 (2005) 2210–2220.
[30] M.V. Mandola, J. Stoehlmacher, W. Zhang, S. Groshen, M.C. Yu, S. Iqbal, H.J. Lenz, R.D. Ladner, A 6 bp polymorphism in the thymidylate synthase gene causes message instability and is associated with decreased intratumoral TS mRNA levels, Pharmacogenetics 14 (2004) 319–327.
[31] M.V. Rockman, G.A. Wray, Abundant raw material for cis-regulatory evolution in humans, Mol. Biol. Evol. 19 (2002) 1991–2004.
[32] P.R. Buckland, B. Hoogendoorn, S.L. Coleman, C.A. Guy, S.K. Smith, M.C. O'Donovan, Strong bias in location of functional promoter polymorphisms, Hum. Mutat. 26 (2005) 214–223.
[33] N.J. Bray, P.R. Buckland, M.J. Owen, M.C. O'Donovan, Cis-acting variation in the expression of a high proportion of genes in human brain, Hum. Genet. 113 (2003) 149–153.
[34] H. Yan, W. Yuan, V.E. Velculescu, B. Vogelstein, K.W. Kinzler, Allelic variation in human gene expression, Science 297 (2002) 1143.
[35] H.H. Yang, Y. Hu, M. Edmonson, K. Buetow, M.P. Lee, Computation method to identify differential allelic gene expression and novel imprinted genes, Bioinformatics 19 (2003) 952–955.
[36] J.C. Knight, B.J. Keating, K.A. Rockett, D.P. Kwiatkowski, In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading, Nat. Genet. 33 (2003) 469–475.
[37] H.S. Lo, Z. Wang, Y. Hu, H.K.H. Buetow, M.P. Lee, Allelic variation in gene expression is common in the human genome, Genome Res. 13 (2003) 1855–1862.
[38] T. Pastinen, R. Sladek, S. Gurd, A. Sammak, B. Ge, P. Lepage, K. Lavergne, A. Villeneuve, T. Gaudin, H. Brandstrom, A. Beck, A. Verner, J. Kingsley, E. Harmsen, D. Labuda, K. Morgan, M.C. Vohl, A.K. Naumova, D. Sinnett, T.J. Hudson, A survey of genetic and epigenetic variation affecting human gene expression, Physiol. Genomics 16 (2004) 184–193.
[39] J. Singer-Sam, J.M. LeBon, A. Dai, A.D. Riggs, A sensitive, quantitative assay for measurement of allele-specific transcripts differing by a single nucleotide, PCR Methods Appl. 1 (1992) 160–163.
[40] J.C. Knight, B.J. Keating, D.P. Kwiatkowski, Allele-specific repression of lymphotoxin-alpha by activated B cell factor-1, Nat. Genet. 36 (2004) 394–399.
[41] D. Hacking, J.C. Knight, K. Rockett, H. Brown, J. Frampton, D.P. Kwiatkowski, J. Hull, I.A. Udalova, Increased in vivo transcription of an IL-8 haplotype associated with respiratory syncytial virus disease-susceptibility, Genes Immun. 5 (2004) 274–282.
[42] X. Liu, M.R. Campbell, G.S. Pittman, E.C. Faulkner, M.A. Watson, D.A. Bell, Expression-based discovery of variation in the human glutathione S-transferase M3 promoter and functional analysis in a glioma cell line using allele-specific chromatin immunoprecipitation, Cancer Res. 65 (2005) 99–104.
[43] N.J. Bray, P.R. Buckland, N.M. Williams, H.J. Williams, N. Norton, M.J. Owen, M.C. O'Donovan, A haplotype implicated in schizophrenia susceptibility is associated with reduced COMT expression in human brain, Am. J. Hum. Genet. 73 (2003) 152–161.
[44] N.J. Bray, L. Jehu, V. Moskvina, J.D. Buxbaum, S. Dracheva, V. Haroutunian, J. Williams, P.R. Buckland, M.J. Owen, M.C. O'Donovan, Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes, Hum. Mol. Genet. 13 (2004) 2885–2892.
[45] N.J. Bray, A. Preece, N.M. Williams, V. Moskvina, P.R. Buckland, M.J. Owen, M.C. O'Donovan, Haplotypes at the dystrobrevin binding protein 1 (DTNBP1) gene locus mediate risk for schizophrenia through reduced DTNBP1 expression, Hum. Mol. Genet. 14 (2005) 1947–1954.
[46] L. Prokunina, M.E. Alarcón-Riquelme, Regulatory SNPs in complex diseases: their identification and functional validation, Expert Rev. Mol. Med. 6 (2004) 1–15.
[47] S. Mottagui-Tabar, M.A. Faghihi, Y. Mizuno, P.G. Engstrom, B. Lenhard, W.W. Wasserman, C. Wahlestedt, Identification of functional SNPs in the 5-prime flanking sequences of human genes, BMC Genomics 6 (2005) 18.
[48] J. Linnell, R. Mott, S. Field, D.P. Kwiatkowski, J. Ragoussis, I.A. Udalova, Quantitative high-throughput analysis of transcription factor binding specificities, Nucleic Acids Res. 32 (2004) e44.
[49] S.L. Coleman, P.R. Buckland, B. Hoogendoorn, C. Guy, K. Smith, M.C. O'Donovan, Experimental analysis of the annotation of promoters in the public database, Hum. Mol. Genet. 11 (2002) 1817–1821.
[50] N.D. Trinklein, S.J.F. Aldred, A.J. Saldanha, R.M. Myers, Identification and functional analysis of human transcriptional promoters, Genome Res. 13 (2003) 308–312.
[51] V.B. Bajic, S.L. Tan, Y. Suzuki, S. Sugano, Promoter prediction analysis on the whole human genome, Nat. Biotechnol. 22 (2004) 1467–1473.
[52] Z. Xuan, F. Zhao, J. Wang, G. Chen, M.Q. Zhang, Genome-wide promoter extraction and analysis in human, mouse, and rat, Genome Biol. 6 (2005) R72.
[53] O.V. Vishnevsky, N.A. Kolchanov, ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters, Nucleic Acids Res. 33 (2005) W417–W422.
[54] R. Gangal, P. Sharma, Human pol II promoter prediction: time series descriptors and machine learning, Nucleic Acids Res. 33 (2005) 1332–1336.
[55] S.L. Coleman, P.R. Buckland, B. Hoogendoorn, C. Guy, K. Smith, M.C. O'Donovan, A streamlined approach to functional analysis of promoter region polymorphisms, BioTechniques 33 (2002) 412–418.
[56] B. Hoogendoorn, S.L. Coleman, C.A. Guy, K. Smith, T. Bowen, P.R. Buckland, M.C. O'Donovan, Functional analysis of human promoter polymorphisms, Hum. Mol. Genet. 12 (2003) 2249–2254.
[57] S.K. Smith, B. Hoogendoorn, C.A. Guy, S.L. Coleman, M.C. O'Donovan, P.R. Buckland, Lack of functional promoter polymorphisms in genes involved in glutamate neurotransmission, Psychiatr. Genet. 13 (2003) 193–199.

[58] C.A. Guy, B. Hoogendoorn, S.K. Smith, S.L. Coleman, M.C. O'Donovan, P.R. Buckland, Promoter polymorphisms in glutathione-S-transferase genes affect transcription, Pharmacogenetics 14 (2004) 45–51.
[59] P.R. Buckland, B. Hoogendoorn, C.A. Guy, S.L. Coleman, S.K. Smith, M.C. O'Donovan, The identification and functional characterization of polymorphisms in the promoters of genes differentially expressed in schizophrenic brains, Biochim. Biophys. Acta 1690 (2004) 238–249.
[60] P.R. Buckland, S.L. Coleman, B. Hoogendoorn, C.A. Guy, S.K. Smith, M.C. O'Donovan, A high proportion of chromosome 21 promoter polymorphisms influence transcriptional activity, Gene Expr. 11 (2004) 233–239.
[61] B. Hoogendoorn, S.L. Coleman, C.A. Guy, S.K. Smith, M.C. O'Donovan, P.R. Buckland, Functional analysis of polymorphisms in the promoter regions of genes on 22q11, Hum. Mutat. 24 (1) (2004) 35–42.
[62] P.R. Buckland, B. Hoogendoorn, C.A. Guy, S.L. Coleman, S.K. Smith, M.C. O'Donovan, An allele of the 5HT2C receptor gene associated with antipsychotic induced weight confers low gene expression, Am. J. Psychiatry 162 (2005) 613–615.
[63] P.C. FitzGerald, A. Shlyakhtenko, A.A. Mir, C. Vinson, Clustering of DNA sequences in human promoters, Genome Res. 14 (2004) 1562–1574.
[64] X. Xie, J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, M. Kellis, Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals, Nature 434 (2005) 338–345.
[65] T. Hirota, I. Ieiri, H. Takane, S. Maegawa, M. Hosokawa, K. Kobayashi, K. Chiba, E. Nanba, M. Oshimura, T. Sato, S. Higuchi, K. Otsubo, Allelic expression imbalance of the human CYP3A4 gene and individual phenotypic status, Hum. Mol. Genet. 13 (2004) 2959–2969.
[66] M. Wjst, Target SNP selection in complex disease association studies, BMC Bioinformatics 5 (2004) 92.
[67] L. Conde, J.M. Vaquerizas, C. Ferrer-Costa, X. de la Cruz, M. Orozco, J. Dopazo, PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes, Nucleic Acids Res. 33 (2005) W501–W505.
[68] S. Mooney, Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis, Brief. Bioinform. 6 (2005) 44–56.
[69] P. Bucher, Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences, J. Mol. Biol. 212 (1990) 563–578.
[70] V.N. Babenko, P.S. Kosarev, O.V. Vishnevsky, V.G. Levitsky, V.V. Basin, A.S. Frolov, Investigating extended regulatory regions of genomic DNA sequences, Bioinformatics 15 (1999) 644–653.
[71] Y. Suzuki, T. Tsunoda, J. Sese, H. Taira, J. Mizushima-Sugano, H. Hata, T. Ota, T. Isogai, T. Tanaka, Y. Nakamura, A. Suyama, Y. Sakaki, S. Morishita, K. Okubo, S. Sugano, Identification and characterization of the potential promoter regions of 1031 kinds of human genes, Genome Res. 11 (2001) 677–684.
[72] N.I. Gershenzon, I.P. Ioshikhes, Synergy of human Pol II core promoter elements revealed by statistical sequence analysis, Bioinformatics 21 (2005) 1295–1300.
[73] M.L. Bulyk, Computational prediction of transcription-factor binding site locations, Genome Biol. 5 (2003) 201.
[74] M.L. Bulyk, P.L. Johnson, G.M. Church, Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, Nucleic Acids Res. 30 (2002) 1255–1261.
[75] S. Mukherjee, M.F. Berger, G. Jona, X.S. Wang, D. Muzzey, M. Snyder, R.A. Young, M.L. Bulyk, Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays, Nat. Genet. 36 (2004) 1331–1339.
[76] A. Kanhere, M. Bansal, Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes, Nucleic Acids Res. 33 (2005) 3165–3175.
[77] A.G. Pedersen, L.J. Jensen, S. Brunak, H.H. Staerfeldt, D.W. Ussery, A DNA structural atlas for *Escherichia coli*, J. Mol. Biol. 299 (2000) 907–930.
[78] R.E. Dickerson, DNA bending: the prevalence of kinkiness and the virtues of normality, Nucleic Acids Res. 26 (1998) 1906–1926.
[79] E.D. Ross, P.R. Hardwidge, L.J. Maher III, HMG proteins and DNA flexibility in transcription activation, Mol. Cell. Biol. 21 (2001) 6598–6605.
[80] J. Nishikawa, M. Amano, Y. Fukue, S. Tanaka, H. Kishi, Y. Hirota, K. Yoda, T. Ohyama, Left-handedly curved DNA regulates accessibility to cis-DNA elements in chromatin, Nucleic Acids Res. 31 (2003) 6651–6662.
[81] J. Kim, S. Klooster, D.J. Shapiro, Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation, J. Biol. Chem. 270 (1995) 1282–1288.
[82] M.J. Packer, M.P. Dauncey, C.A. Hunter, Sequence-dependent DNA structure: dinucleotide conformational maps, J. Mol. Biol. 295 (2000) 71–83.
[83] D. Nathan, D.M. Crothers, Bending and flexibility of methylated and unmethylated *Eco*RI DNA, J. Mol. Biol. 316 (2002) 7–17.
[84] C. Rivetti, M. Guthold, C. Bustamante, Wrapping of DNA around the *E. coli* RNA polymerase open promoter complex, EMBO J. 18 (1999) 4464–4475.
[85] J.D. Parvin, R.J. McCormick, P.A. Sharp, D.E. Fisher, Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor, Nature 373 (1995) 724–727.
[86] P. Konig, T.J. Richmond, The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility, J. Mol. Biol. 233 (1993) 139–154.
[87] A.K. Nagaich, E. Appella, R.E. Harrington, DNA bending is essential for the site-specific recognition of DNA response elements by the DNA binding domain of the tumor suppressor protein p53, J. Biol. Chem. 272 (1997) 14842–14849.
[88] P. Shore, A.D. Sharrocks, The MADS-box transcription factors, Eur. J. Biochem. 229 (1995) 1–13.
[89] T.B. Acton, J. Mead, A.M. Steiner, A.K. Vershon, Scanning mutagenesis of Mcm1: residues required for DNA binding, DNA bending and transcriptional activation by a MADS-box protein, Mol. Cell. Biol. 20 (2000) 1–11.
[90] J. Mead, A.R. Bruning, M.K. Gill, A.M. Steiner, T.B. Acton, A.K. Vershon, Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast, Mol. Cell. Biol. 22 (2002) 4607–4621.
[91] T. Acton, H. Zhong, A.K. Vershon, DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein, Mol. Cell. Biol. 17 (1997) 1881–1889.
[92] E.A. Carr, J. Mead, A.K. Vershon, Alpha1-induced DNA bending is required for transcriptional activation by the Mcm1–alpha1 complex, Nucleic Acids Res. 32 (2004) 2298–2305.
[93] E.C. Murphy, V.B. Zhurkin, J.M. Louis, G. Cornilescu, G.M. Clore, Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation, J. Mol. Biol. 312 (2001) 481–499.
[94] L. Tsai, L. Luo, Z. Sun, Sequence-dependent flexibility in promoter sequences, J. Biomol. Struct. Dyn. 20 (2002) 127–134.
[95] D. Nègre, C. Bonod-Bidaud, C. Oudot, J.F. Prost, A. Kolb, A. Ishihama, A.J. Cozzone, J.C. Cortay, DNA flexibility of the UP element is a major determinant for transcriptional activation at the *Escherichia coli* acetate promoter, Nucleic Acids Res. 25 (1997) 713–718.
[96] D.B. Nikolov, H. Chen, E.D. Halay, A. Hoffman, R.G. Roeder, S.K. Burley, Crystal structure of a human TATA box-binding protein/TATA element complex, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 4862–4867.
[97] S.C. Schultz, G.C. Shields, T.A. Steitz, Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees, Science 253 (1991) 1001–1007.
[98] E.D. Ross, A.M. Keating, L.J. Maher III, DNA constraints on transcription activation in vitro, J. Mol. Biol. 297 (2000) 321–334.
[99] K.J. Breslauer, R. Frank, H. Blocker, L.A. Marky, Predicting DNA duplex stability from the base sequence, Proc. Natl. Acad. Sci. U. S. A. 83 (1986) 3746–3750.